Fine-Tuning a Sentiment Analysis Model

Using a Pre-trained Model from Hugging Face

Problem Statement:

Businesses receive large amounts of customer feedback. They can receive feedback in the form of surveys, social media comments, and reviews left on their website. To use this data, businesses need to analyze the data. For this project, I propose creating a model to predict the sentiment of customer feedback. The model will be created using a pre-trained hugging face model that will be trained on a dataset of product reviews for an online clothing retailer. The goal of this model will be to help businesses use customer feedback data to inform decisions about product development.

Data Wrangling:

The dataset for this project is product reviews from an online clothing retailer. There are ten features in the dataset. The review text and rating feature were kept for this project. The remaining features were removed. The title feature was removed because not all review texts had a title. The other features were removed because they were not needed for the model training. Below is a preview of the dataset before it was cleaned.

	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Initmates	Intimate	Intimates
1	1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants
4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses

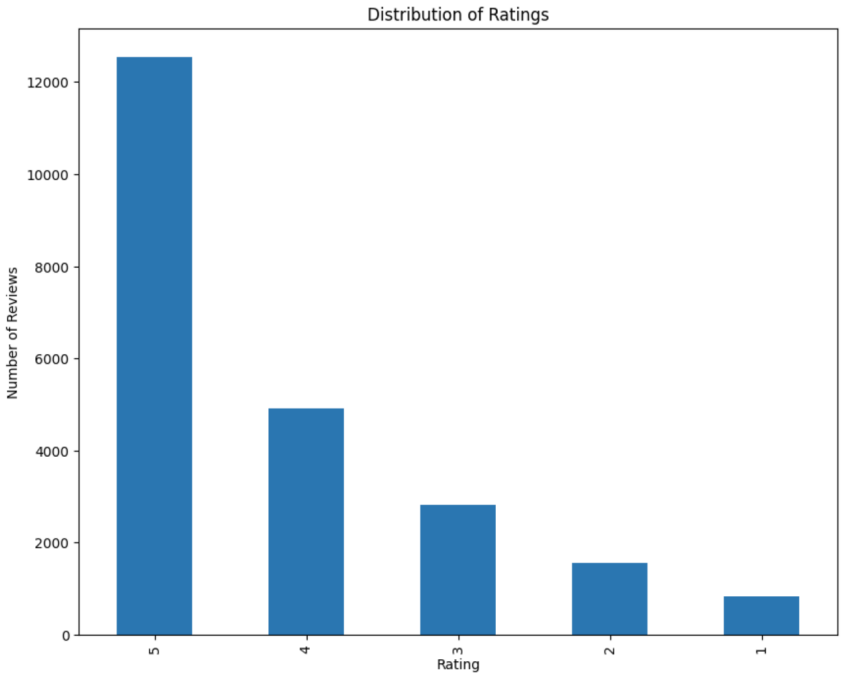
The cleaned dataset contains the review text as type object and the rating as an integer. The rating is a scale of 1-5 with 1 being negative and 5 being positive. There were about 4% review texts with null values and none in the ratings. The instances with null values were removed, leaving a dataset with no missing values in either feature.

The dataset is left with 22,641 entries and two columns. A snapshot of the dataset can be viewed below.

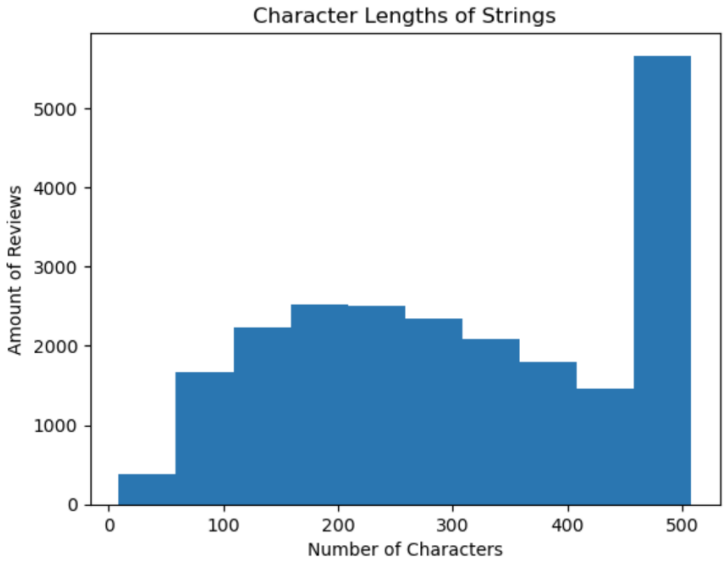
	Review Text	Rating
0	Absolutely wonderful - silky and sexy and comf...	4
1	Love this dress! it's sooo pretty. i happene...	5
2	I had such high hopes for this dress and reall...	3
3	I love, love, love this jumpsuit. it's fun, fl...	5
4	This shirt is very flattering to all due to th...	5

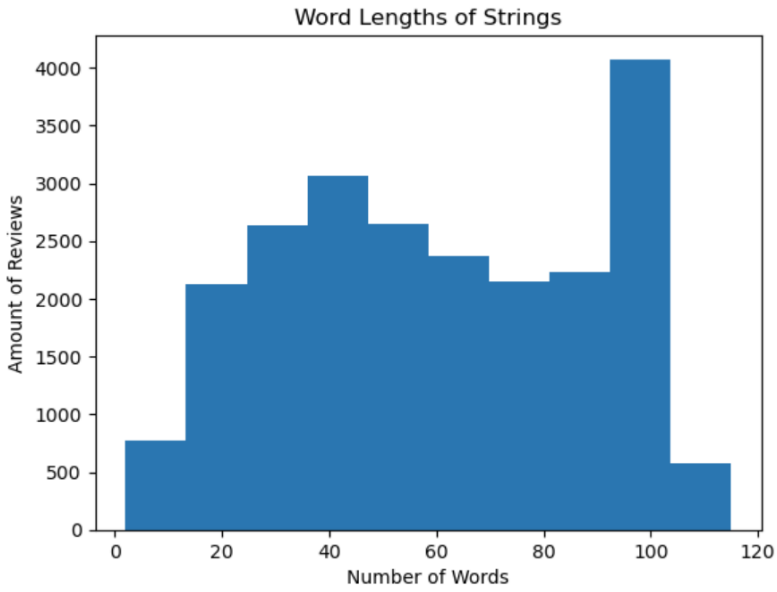
Exploratory Data Analysis and Initial Findings:

First, I took a closer look at the rating feature. This is an integer with values 1, 2, 3, 4, and 5. The scale ranges from positive to negative with 5 representing positive reviews and 1 representing negative reviews. The average of the ratings is about 4 which can lead us to assume that there are more higher-rated, positive reviews in the dataset. Below is a distribution of the ratings. Here it can be seen that there are mostly positive reviews with a rating of 5. There are fewer reviews with each lower rating.



Next, I examined the review text. The review text was converted to a string. Histogram plots were used to display the character and word length of the review text strings. In the first histogram, it can be seen that the character length ranges from 0-500 characters. The word count for each review ranges from 0-120.



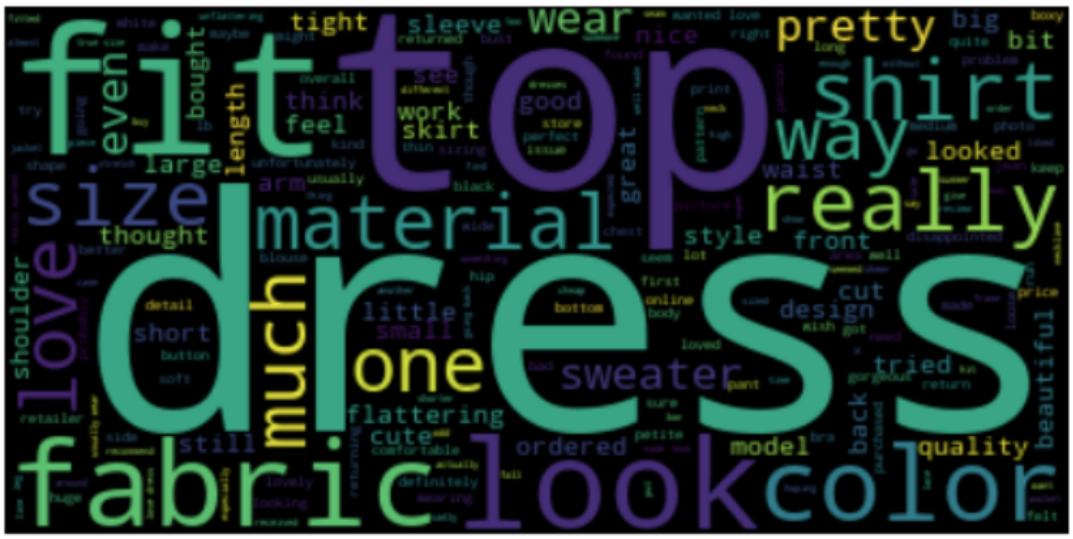


Since words will be the focus of the sentiment analysis, I wanted to see which words occurred the most frequently in the dataset based on rating. The data was further cleaned by removing punctuation, contractions and stopwords. The data was converted to all lowercase and the words were lemmatized. The review text strings were then

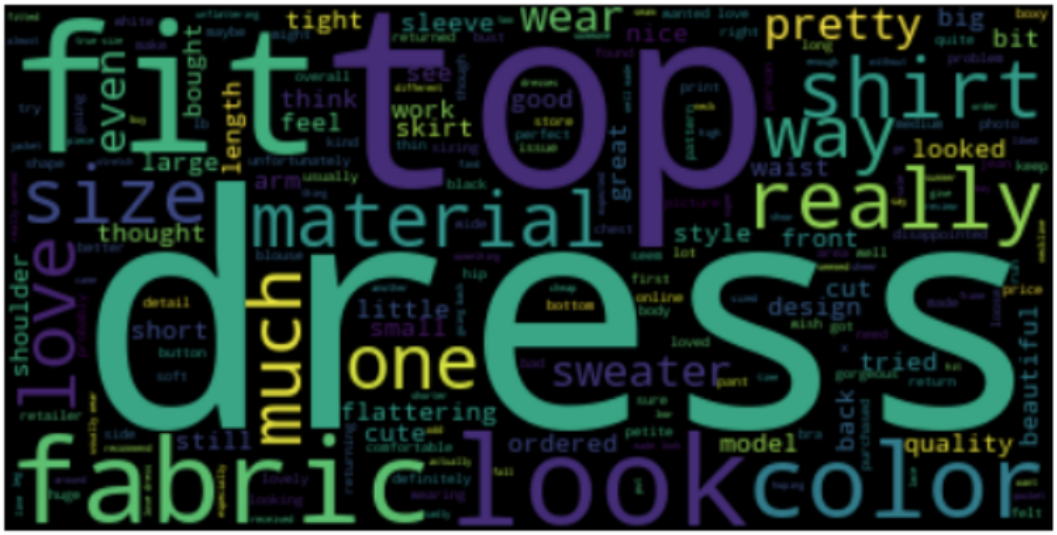
tokenized. To view differences between the numeric ratings, the tokenized strings were divided into positive (rating 5 or 4), neutral (rating 3), and negative (rating 2 or 1) sets. The most frequent words were found using FreqDist(). FreqDist() gives a list of words and their number of occurrences in the dataset. A word cloud was created for each set to give a visual representation of the common words in the dataset. The word clouds can be seen below.

Rating 5 and 4

Rating 3



Rating 2 and 1



Looking at the word cloud, there are some common words between the sets. These words include dress, top, love, look, and fabric. There are differences in the occurrence of these words between the sets.

The Model:

The distilbert-base-uncased model from Hugging Face was used in this project. This model was chosen because it is a smaller and faster version of the BERT model. The model was trained on the clothing reviews dataset to create a sentiment analysis model. To train the dataset, the model was split into test, train, and validation datasets. Splitting the datasets left about 13,500 entries for training and about 4,500 entries for the test and validation dataset. The distilbert-base-uncased tokenizer was used to prepare the dataset for training. The model was set up to look for 5 class labels. The class labels would be the rating 1-5. The default training parameters were used for the training arguments.

Below are the evaluation metrics for the trained model.

Evaluation Metrics	**Value**
Loss	0.928
Accuracy	0.672
F1 Score	0.668
Runtime	19.640

Ideas for Further Research:

One of the problems in the dataset was the uneven distribution of ratings. There were more 5 ratings than any other rating. To improve the model's accuracy, I would include more data for the remaining rating values. The model is trained only on clothing reviews. To expand the adaptability of the model for sentiment analysis, I would include reviews and ratings on a variety of products. The other factor to consider is the low accuracy of the model. In the future, I would test different model parameters to improve the model's performance.